『외국어교육』 제17집, 2015. 08 (91-112) ISSN 2233-6095

외국어교육을 위한 음성 합성의 원리와 연결 음성 합성

정 현 성 (한국교원대)

한국교원대학교
외국어교육연구소

2015. 08.

외국어교육을 위한 음성 합성의 원리와 연결 음성 합성

정 현 성 (한국교원대)

I. 들어가는 말

음성 합성은 기본적으로 컴퓨터나 기계가 인간의 음성을 합성해 발화하도록 하는 기술이다. 음성 합성은 사람의 육성을 직접 사용하지 않고 자동으로 책을 읽어주는 프로그램이나 기차역 같은 곳에서 흘러나오는 장내 방송 등을 통해 접할 수 있을 정도로 의외로 우리의 생활에 밀접하게 다가와 있다. 음성 합성은 단순히 정해진 문장만을 읽는 것에서 우리말을 외국어로, 외국어를 우리말로 자동 통역하는 과정에서도 사용할 수 있는 기술로 음성 합성의 기본 원리가 그렇게 복잡하지는 않다. 그러나 오늘날 사람이 말하는 것과 같이 자연스럽게 합성음을 구현하는 것은 쉽지는 않다. 본 논문에서는 음성합성의 기본 원리에 대해 살펴보고, 음성 합성을 쉽게 이해하고 적용해 볼 수 있는 연결 음성 합성을 소개한다.

II. 음성 합성의 기본 원리

기차역에서 가장 많이 들을 수 있는 것은 “The train is bound for 00.”라는 장내 방송이다. 이러한 장래 방송을 하기 위해서 일일이 모든 문장을 녹음하는 것도 한 가지 방법이다. 보다 적은 노력으로 합성한다면 “00”로 표시된 부분은 따로 녹음해 음성 자료로 만들어 두고, 임의로 역명을 정해 위에 해당하는 하나의 문장만을 녹음한 후 00에 해당하는 항목은 자판에 입력하면 녹음되어 있는 음성 자료에서 자동으로 음성을 불러드리도록 하는 방법을 사용할 수 있다. 기본 문장 외에 더 필요한 음성 자료는 목적지를 나타내기 위한 “Pittsburgh, Philadelphia, Newark, New York” 등의 역명이다. 이렇게 함으로써 모든 문장을 일일이 녹음할 필요가 없기 때문에 음성 자료를 저장하기 위한 기억 공간도 많이 필요하지 않다는 장점이 있다. 이 과정을 통해 구축된 역명 음성자료는 다른 내용의 공지사항이 있더라도 기본적인 하나의 문장 하나만 생성하면 역명과 시각에 따라 일일이 다시 녹음할 필요 없이 간단히 대치가 가능하다. 예로 제시된 음성 합성의 과정을 그림으로 나타내면 다음과 같다.
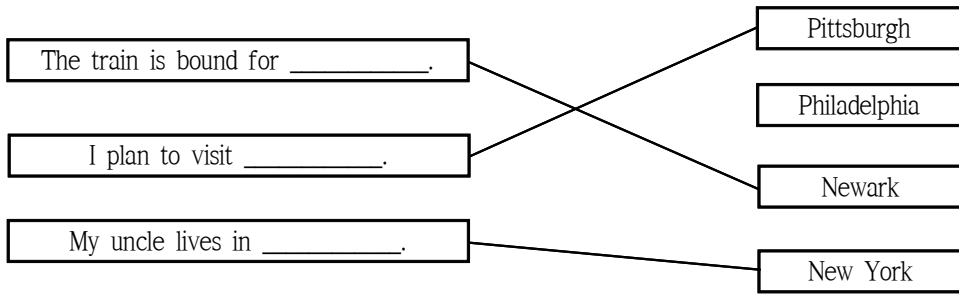
그림 1. 어구 단위 음성 합성의 예

위에 제시된 예에서는 음성 합성의 단위를 어절 또는 어구, 개별 단어 등으로 비교적 큰 단위를 사용하고 있지만, 몇 가지 유형에 따른 문장이 정해져 있지 않고, 결과물로 생성해야 할 문장이 수 만개라면 새로운 어절이나 어구가 등장할 때 마다 일일이 녹음할 수 없기 때문에 다른 방법을 모색해야한다. 따라서 음성 자료의 양과 노력을 최소화 하면서 자연스러운 합성음을 생성하기 위해 다양한 기술이 개발되고 있다. 음성합성은 크게 인간의 조음 기관과 발성 원리를 구현해 사람의 목소리를 직접 활용하지 않고 순수하게 인공적인 합성음을 생성하는 방법과 사람의 목소리를 직접 활용하여 녹음된 음파를 저장했다가 음성 합성에 이용하는 방법으로 구분할 수 있다. 전자에 속하는 음성 합성 기법은 규칙 기반 포먼트 합성 rule-based formant synthesis과 조음 기반 음성 합성 articulatory synthesis이 해당되고, 후자의 경우에는 코퍼스 기반 연결 음성 합성 concatenative synthesis과 HMM 기반 음성 합성이 속한다. 본 논문에서는 이 중 코퍼스 기반 연결 음성 합성 concatenative synthesis에 대한 기본 원리와 응용에 대해 살펴보고자 한다.

III. 음성 합성의 기술

음성합성은 크게 인간의 조음 기관과 발성 원리를 구현해 사람의 목소리를 직접 활용하지 않고 순수하게 인공적인 합성음을 생성하는 방법과 사람의 목소리를 직접 활용하여 녹음된 음파를 저장했다가 음성 합성에 이용하는 방법으로 구분할 수 있다. 전자에 속하는 음성 합성 기법은 규칙 기반 포먼트 합성과 조음 기반 음성 합성이 해당되고, 후자의 경우에는 코퍼스 기반 연결 음성 합성과 HMM 기반 음성 합성이 속한다. 이 장에서는 이러한 합성 방식에 대한 기본 원리를 살펴보고 음향 음성학적으로 활용과 구현이 비교적 쉬운 코퍼스기반 연결 음성 합성에 대해 자세히 살펴보고자 한다.

1. 규칙 기반 포먼트 합성

규칙 기반 포먼트 합성 기법이 등장하기 이전에 처음으로 체계화된 음성 합성기의 모습을 갖춘 것은 1939년 뉴욕 세계박람회에서 선보인 'Voder(Dudley et al., 1939)' 합성기였다.
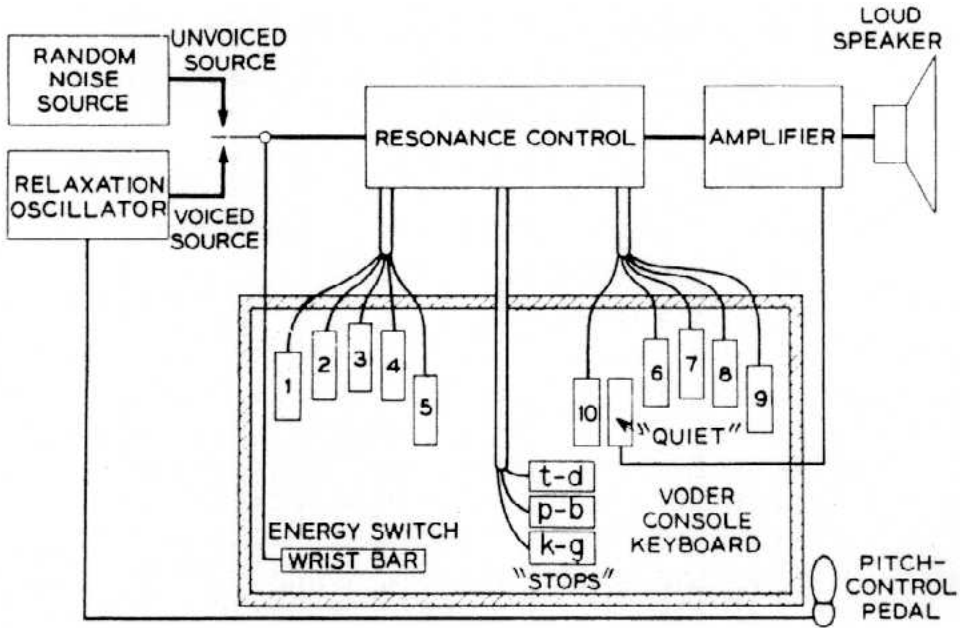
그림 2. Voder 음성 합성기(Dudley et al., 1939; Klatt(1987)에서 재인용)

Voder 합성기는 피아노 건반을 두드리는 것처럼 인간이 직접 손과 발을 사용해 조작하도록 설계되었다. 유/무성 음원 voicing source과 소음 음원 noise source을 통제하는 손목봉 wrist bar, 성대 진동의 기본 주파수 fundamental frequency를 조절하는 밟기판 foot pedal으로 구성되어 있다. 음원은 10개의 대역 통과 전자 필터 bandpass electronic filters를 거치면서 개별 소리를 생성하도록 고안되었고 출력의 크기는 사람이 직접 조작하도록 설계되었다. 문장 합성음을 생성하기 위해서는 상당한 노력과 기술이 필요했고, 합성음의 이해가능성 intelligibility는 미미했지만 음성 합성의 가능성을 열어준 모델이었다.

이어서 등장한 것은 해스킨스 연구소 Haskins Laboratory에서 개발한 '유형 재생 pattern playback' 합성기였다(Cooper et al. 1951).
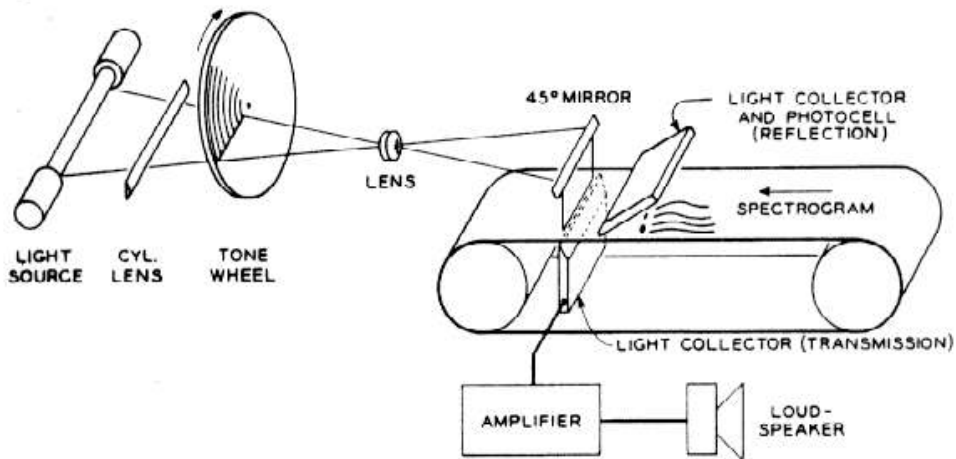
그림 3. 유형 재생 합성기(Cooper et al., 1951; Klatt(1987)에서 재인용)

이 합성기는 120 Hz 간극의 고조파 harmonics 진폭 amplitude을 광학적으로 조작해 스펙트로그램 정보가 담긴 OHP 용지에 통과시켜 소리를 생성하는 모델이다. 광학적 에너지가 광원 light source에서 출발해 실린더 렌즈 cyl. lens를 거쳐 음조 조절 바퀴 tone wheel에서 120 Hz 간극의 고조파가 생성되고, 렌즈를 통과한 이 고조파는 거울을 거쳐 스펙트로그램 정보가 칠해진 움직이는 OHP 용지에 투과되면서 시간축에 따른 소리를 생성하는 원리이다. 이 모델은 주로 음절 단위의 합성음을 생성해 다양한 음성을 변별하는 하는데 사용되었다.

Voder와 유형 재생 합성기가 스펙트로그램의 유형을 활용해 합성을 시도한 것이라면 규칙 기반 포먼트 합성 기법은 음성이 생성되는 음향 이론을 바탕으로 소스-필터 모델 source-filter model을 충실히 반영한 합성 방식으로 규칙에 의해 개별음을 직접 생성하고 조합하는 방식이다. 이 방식에서는 음성파의 생성 기구를 음원의 생성과 성도의 형태에 의한 조음(필터 특성)으로 나누어 규칙에 따라 음원과 성도 특성을 제어함으로 합성음을 생성하였다.
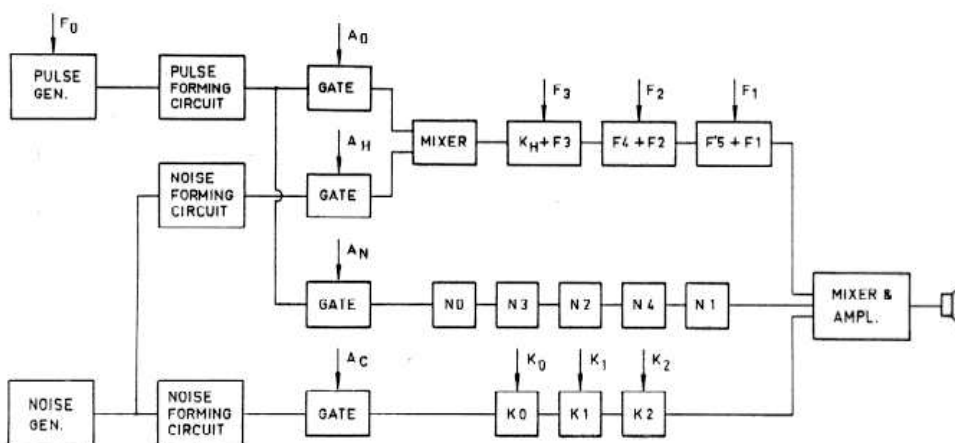
그림 4. Orator Verbis Electris II (OVE II; Fant & Martony 1962; Klatt(1987)에서 재인용)

위 그림은 초기 포먼트 합성기 중의 하나인 Orator Verbis Electris(OVE I; Fant 1953)가 진화한 OVE II의 계통도이다. 크게 세 영역의 회로로 분화되어 있는데 가장 상단의 ‘KH+F3’부터 ‘F5+F1’까지는 모음, 중간의 ‘N0’부터 ‘N1’까지는 비음, 하단의 ‘K0’부터 ‘K2’까지는 장애음의 성도 전이 기능 vocal tract transfer function을 구현한 것이다. 음원 sound source의 경우 상단의 성대 진동 발성부 pulse gen.에서 시작되는 성대 진동 형성 회로 pulse forming circuit는 유/무성 voicing을 통제하고, 소음 발성부 noise gen.의 경우 중간의 소음 형성 회로 noise forming circuit는 기식 aspiration 소음을, 하단의 소음 형성 회로는 마찰 frication 소음을 통제한다.

실제로 이러한 회로를 통제하고 소리를 생성하기 위해서는 각 부분을 적절하게 통제할 수 있는 복잡한 규칙이 필요하다. 그리고 직접 회로를 작동해 소리를 생성하기 보다는 컴퓨터를 활용해 회로의 기능을 대치하게 된다. 이러한 규칙을 바탕으로 최초로 음소를 합성한 것은 Kelly와 Gerstman(1961; 1962)이 최초였고, 이후 합성의 효율성과 자연성을 향상시키기 위해 음원의 생성 방식은 어떻게 할 것인지, 음원과 성도 전이 기능을 직렬식 cascade으로 연결할 것인지, 병렬식 parallel으로 연결할 것인지 등을 두고 연구를 거듭하면서 Holmes(1973)의 병렬식 포먼트 합성기, Klattalk(Klatt 1980), MITalk(Allen et al. 1987) 등으로 발전해 갔다. 이러한 포먼트 합성의 경우 처음에 음소 단위의 소리만 생성하고 운율의 자연발화를 복사해 사용했지만 Mattingly(1966) 이 후 운율을 직접 생성하기 위한 노력이 시작되었다.

2. 조음기반 음성 합성

인간의 조음 기관을 직접적으로 모델링하여 합성음을 생성하는 기술이다. 규칙 기반 포먼

트 합성 기술이 인간의 조음적 특성을 직접적으로 반영하고 있지 못하기 때문에 좀 더 실질적으로 조음 기관의 움직임을 반영해 합성음을 생성하고자 했다. 후두에서 입술에 이르기까지의 성도 단면을 단순화 해 개별음에 필요한 조음 기관별 영역별 함수 값을 수학적 계산에 따라 도출해 내어 영역별 값을 선형보간법 linear interpolation에 따라 연결해 합성음을 생성하는 방식이다.
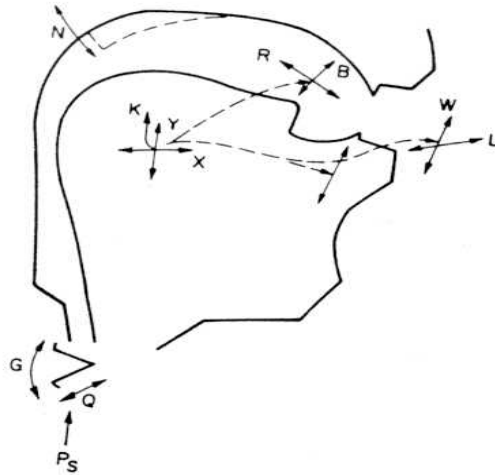
그림 5. 조음 기반 음성 합성에 사용된 성도 모형(Coker 1976; Klatt(1987)에서 재인용)

위 그림은 Coker(1976)에서 제시된 성도 단면도인데, 입술의 움직임에 따라 W, L, 혓몸의 움직임에 따라 Y, X, K, 혀끝의 움직임에 따라 B, R, 비강의 개방에 대해서는 N, 음원에 대해서는 P, G, Q 영역의 함수를 수학적으로 계산하여 합성음을 생성하는 원리이다.

이론적으로 가장 자연음에 가까운 합성음을 생성할 수 있겠지만, 조음 기관의 특성을 반영해 수학적 계산을 가능하게 해 줄 만큼 조음 자료의 축적이 충분치 못하여 최적화된 합성음 품질을 생성하지는 못하고 있다.

3. HMM(Hidden Markov Model) 기반 음성 합성

HMM 기반 합성의 경우 개별 소리의 음향 특성 spectral information과 환경별 변화를 반영한 여기 매개변수 excitation parameter를 통계적인 음향 모델로 구현하여 합성하는 기법이다. 이 기법은 적은 음성 자료로도 어느 정도 양호한 음질의 합성음을 생성하는 것이 가능하고, 다양한 환경에 따른 음성과 운율의 변화가 상대적으로 용이하다는 장점이 있다. 즉 파형을 직접적으로 활용하지 않고 사람의 목소리를 직접 녹음한 적은 량의 음성 자료를 환경별로

분석한다. 분석에 따른 음파에 대한 통계적 정보를 매개변수의 형태로 축적한 다음 새로운 문장이 입력되면 이러한 매개변수를 활용하여 새로운 문장의 환경에 가장 근접한 합성음을 생성하는 방식이다. 매개변수는 음원을 검출하는 매개변수, 소리의 스펙트럼 특성을 검출하는 매개변수가 있다. 이와 더불어 기본주파수 F0와 합성 단위 길이 예측 모델이 함께 수반된다. 합성 단위의 스펙트럼 특성을 검출한 매개변수는 하나의 음소에 대해 하나의 특성만 가지고 있는 것이 아니고 다양한 환경에 따른 특성을 반영하고 있다. 그렇기 때문에 합성을 할 경우에는 새로운 환경에 가장 적합한 매개변수 정보를 제공해 자연스러운 합성음 생성이 가능하다.

합성 과정은 훈련부와 합성부로 구분되는데 훈련부는 다양한 음성 자료를 통해 매개변수를 검출하는 단계이다. 훈련을 위해 사용되는 음성 자료에는 음성자료가 낭독 발화인지, 자유 발화인지, 문장의 구문 및 음운·운율 구조, 단어의 품사 등 다양한 환경별 정보에 대한 표지가 있어야 하고 이러한 표지를 바탕으로 훈련용 음성 자료에 대한 다양한 통계적 매개변수가 구축된다. 합성부에서는 새로운 문장이 입력되면 훈련부에서 활용한 것과 똑같은 표지를 새로운 문장에 적용하여 자동으로 분석하고 이렇게 분석된 자료에 대해 훈련부에서 이미 구축된 다양한 매개변수 중에 가장 근접한 것을 도출하여 합성음을 생성하게 된다. 따라서 훈련부에서 활용되는 문장의 양이 다양할수록 더욱 자연스러운 합성이 가능하게 된다.

HMM 기반 음성 합성은 확장성이 크고 다양한 환경에 대해 비교적 균일한 음질의 합성음을 생성할 수 있다는 장점이 있는 반면, 매개변수의 검출 과정과 결과가 지나치게 컴퓨터에 의존적이라는 점에서 음성·언어학적 접근이 쉽지 않다.

4. 연결 음성 합성

1990년대 이후에 합성음의 자연성을 향상시키기 위해서 사람의 목소리를 미리 녹음해 두고 적절한 합성 단위로 저장해 두었다가 필요한 합성 단위만을 골라 음성 합성에 활용하는 연결 음성 합성이 자연성의 측면에서 효과적이고 합성을 위한 계산 속도가 빠르다는 점에서 음성 합성의 주류를 이루었다. 이 방법의 가장 큰 고민은 합성의 기본 단위를 무엇으로 하고, 그 단위의 연결을 어떤 방식으로 할 것인가 하는 것이다. 이러한 고민에 따라 음성 합성의 기본 단위를 음소, 반음소 diphone, 반음절 demi-syllable, 음절 syllable, 모음-자음열-모음(VCV) 또는 자음열-모음-자음열(CVC) 등으로 다양하게 사용해 왔다. 이러한 기본 단위의 수는 개별 언어마다 다르기 때문에 각 언어의 특성을 고려하여 기본 단위를 선택하여야 한다.

반면에 가장 많이 쓰는 문장 단위 이상의 대규모 음성 자료를 구축하는 기술도 필요하다. 다양한 음향적 정보를 표시한 후 주어진 상화에 가장 적절한 합성 단위를 고르는 기술을 unit selection synthesis라고 한다. 합성 단위의 크기와 종류에 따라 합성음의 음질에 차이가 있으며 자연스럽고 높은 음질의 합성음을 생성하기 위해서는 합성 단위로 사용되는 음성 자료의

크기가 비교적 대용량이어야 한다. 최근에는 기본 단위를 보다 탄력적으로 생성해 낼 수 있는 환경 지향 클러스터링 COC: context-oriented clustering 기법이 개발되어 대규모 음성 자료에서 합성음을 자연스럽게 실현시키기에 가장 적절한 기본 단위를 자동 검출하게 되었다. 이미 구축된 음성 자료를 사용하게 되면 다양한 환경에 따른 음성과 운율의 변화가 쉽지 않다는 단점이 있다.

사람의 목소리를 음원으로 할 경우, 휴대 통신 기기처럼 적은 용량으로 음성합성을 실현할 것인가, 아니면 속도나 저장 용량에 크게 제한 없이 음성합성을 실현한 것인가에 따라 음성합성의 접근 방법이 달라진다. 적은 저장 용량과 휴대성을 지향할 경우, 모든 음성 환경을 반영해 실현이 가능한 모든 음절을 녹음한 후, 음성 합성의 목적과 지향성에 맞게 다양한 합성단위로 분석해 자료로 구축한다. 음성 합성을 위해 문자열을 입력하면 그 문자열에 적절한 합성단위를 음성 자료에서 찾아 합성음을 생성한다. 새롭게 제시되는 모든 문자열을 녹음하는 것이 아니기 때문에 저장 용량과 처리 속도를 줄일 수 있는 이점이 있다.

개별 음소를 기본 단위로 사용할 경우 음소의 수가 한정되어 있기 때문에 음성 합성에 사용할 기본 단위의 개수가 가장 적고 저장 공간이 많이 필요하지 않으며, 합성을 위해 필요한 계산 시간도 짧은 이점이 있다. 하지만 두 개의 다른 음소는 연결부의 음성 파형이 다르기 때문에 동시조음 coarticulation을 반영한 합성을 위해 연결부를 자연스럽게 이어줄 수 없다는 단점이 있다. 또, 음소 단위만으로는 음운 환경에 따른 다양한 변이음을 반영하기 어려워 자연스러운 합성음을 생성하기 위해서는 음의 변이와 운율의 변화를 위해 여러 가지 음운 변형 규칙이 필요하다는 한계를 지닌다고 볼 수 있다. 동시조음의 불일치를 극복하기 위해 등장한 단위가 반음소와 반음절이다. 반음소는 연결된 두음소열에서 앞 음소의 중간에서 시작해 뒷 음소의 중간에 끝나는 음향 신호를 잘라 합성의 기본 단위로 사용한다. 이렇게 함으로써 서로 다른 두 음소 간의 음향 신호 간의 불일치와 동시조음의 문제를 자연스럽게 극복할 수 있다. 예를 들어 “mother”의 경우 아래 방식으로 절단하여 기본 단위로 사용한다.

그림 6. 반음소 diphone의 추출 원리

첫 번째 열에 제시된 합성 단위가 음소라면 점선으로 표시된 네모 상자는 음소의 연속 구간을 앞 음소와 다음 음소의 중간에서 잘라 생성한 반음소이다. 예를 들면 위에서 생성된 “#-m”와 “m-ʌ” 반음소는 “month”라는 단어의 합성음을 생성할 때도 사용할 수 있고, “ð-ə,” “ə-r,” “r-#” 반음소는 “father”를 합성할 때도 사용할 수 있기 때문에 굳이 새로운 합성음을 생성하기 위해서 “month”나 “father”를 다시 녹음할 필요가 없어진다. 반음소를 합성 단위로

사용할 경우 절단 부분은 개별 음소의 안정 구간에서 시작되고 변화하는 부분은 이미 반음소 안에 포함되어 있기 때문에 개별 음소가 결합하면서 생길 수 있는 부자연스러움을 해소할 수 있다. 즉 결합되는 부분은 같은 음소의 절단면이 결합하는 것이기 때문에 자연스러운 합성음을 생성하는 것이 개별 음소를 단위로 해서 합성할 때 보다 더 용이해진다.

이와 유사한 단위로는 모음 사이의 자음은 모두 합성 단위에 포함시키는 VCV 단위를 사용하면 모음의 안정 구간이 결합되면서 자음 간의 동시조음의 특성을 더 잘 반영할 수 있다. 즉 모음 사이에 오는 자음열을 하나의 단위로 합성하는 방식이다. 예를 들어 영어의 "aspirin"을 합성할 경우 반음소를 단위로 했을 때는 "#-æ," "æ-s," "s-p," "p-r," "r-ɪ," "ɪ-n," "n-#"등 7개의 합성 단위가 필요하지만 모음 중간 부분만을 절단해 사용하는 VCV 단위를 합성단위로 사용하면 "#-æ," "æ-s-p-r-ɪ," "ɪ-n-#" 등 3개의 합성 단위만 사용하면 되기 때문에 합성 단위가 많아지면서 생기는 부자연스러움을 극복할 수 있다. 이것은 동시조음을 반음소일 때 보다 더 잘 반영할 수 있지만 축적해야 할 합성 단위가 반음소일 때 보다 많아지기 때문에 음성 자료의 저장 공간이 커진다는 단점이 있다. 하지만 컴퓨터 기기의 용량과 처리 속도가 빨라지면서 이것은 크게 문제가 되지는 않는다. 그럼에도 불구하고 여전히 연결 부분의 자연성을 위해서 원 음성 자료는 운율을 배제한 상태에서 녹음하기 때문에, 자연스러운 운율을 생성하기 위해서는 환경과 발화의 종류에 따른 여러 가지 규칙이 필요하다.

IV. 문자-음성 변환시스템 TTS: Text-to-Speech

일반적으로 연결 음성 합성을 통해 문자를 음성으로 변환하는 것은 문자-음성 변환시스템 TTS: Text-to-Speech이라고도 하는데, 현재 실생활에서 비교적 일반적으로 쓰이고 있는 기술이다. 이 문자-음성 변환시스템의 합성 과정은 <그림 7>과 같다.
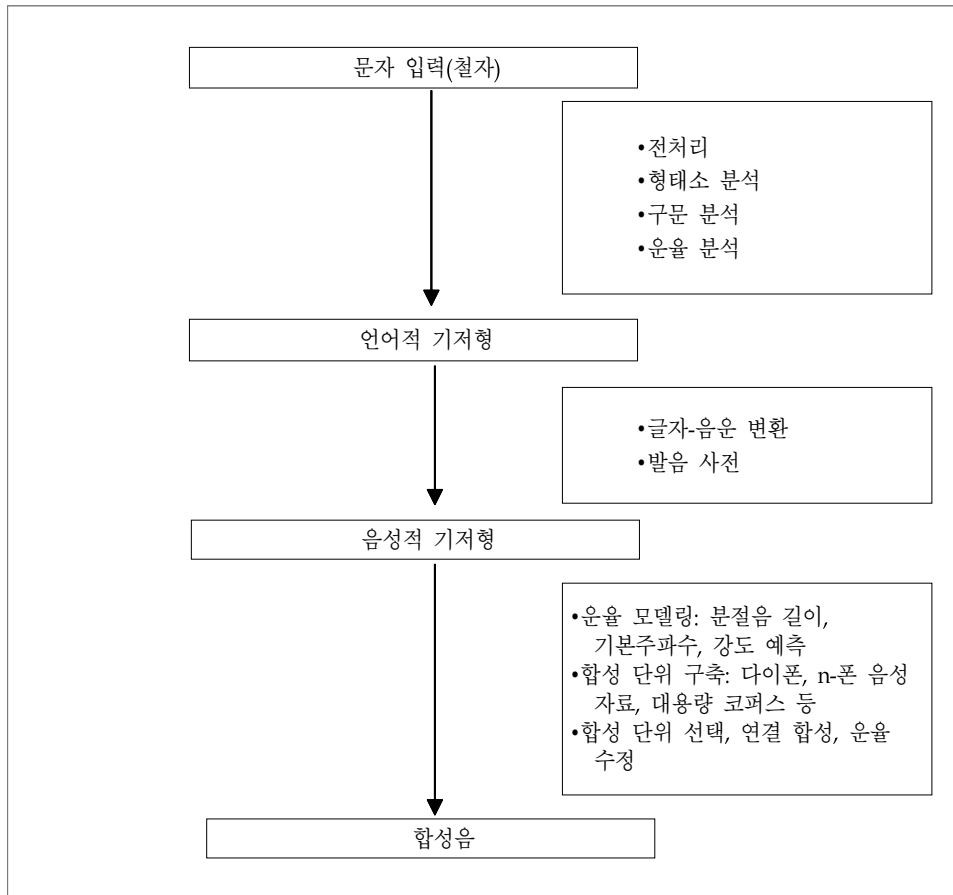
그림 7. 문자-음성 변환시스템의 합성 과정

위 단계에서 알 수 있듯이 전처리부에서는 다양한 형태의 문자열을 동일한 문자열로 생성해 주기 위한 과정이다. 즉, 기호나 약어, 특수 문자, 수학표기 등을 있는 그대로가 아니고 아래 보기와 같이 문자열로 바꾸어 주는 과정이다.

5 km → five kilometers
123-4567 → one two three four five six seven

전처리를 위해서는 이러한 기호들을 일반 문자열로 변환할 때 참조할 수 있는 별도의 기호/약어 사전을 구축할 필요가 있다

문장음성 변환시스템에서 구분석기는 자연스러운 운율 정보(억양, 지속 시간 등)를 생성하기 위해 문장의 구문 정보를 추출해 내는 역할을 한다. 자연어 처리에서 말하는 파서(parser)

와 비슷한 역할을 하지만 실시간 처리를 요구하는 문장 음성 변환시스템의 성격 때문에 주로 구단위로 파싱을 한다.

운율 분석은 입력된 문장의 의미, 화용 정보를 반영하여 구단위로 파싱된 정보를 실제 의미 단위로 분석하는 과정이다. 이것은 문법 단위가 항상 운율 단위와 일치하지는 않는 현실을 반영한 과정이다.

그 다음 과정은 생성된 문자열을 음성 합성 시스템이 올바르게 읽을 수 있도록 소리 나는 대로 전사한 발음열로 바꿔주는 글자-음운 변환이다. 영어 음성 합성 시스템의 경우 영어의 모음과 자음 및 후어휘규칙 등에 의해 단어 내나 단어 간에 적용되는 음성, 음운규칙을 활용해 기준이 되는 영어 억양에 맞추어 글자를 발음열로 바꾸어 준다(예: this year /ðɪs jɪɹ/ → [ðɪʃɚ]). 이 발음열은 글자-음운 변환 규칙을 적용하거나 예외발음 사전 등을 활용하여 생성할 수 있다.

그 다음 과정은 이렇게 생성된 발음열을 실제 음성 파형으로 전환하는 과정이다. 실제 음성 파형으로 전환하기 위해서는 합성 단위를 무엇으로 할 것인가가 우선 결정되어야 하고, 음성 파형에 구현될 분절음의 길이, 기본주파수 값, 강도 등의 운율 정보가 있어야 한다. COC 기법에서처럼 합성 단위가 대단위 코퍼스를 통해 단어나 단어 보다 큰 단위를 사용하는 연결 합성이라면 운율 정보는 환경에 맞는 합성 단위만 추출하면 자연스럽게 부차적으로 따라오고 극히 제한적으로만 운율을 예측하기만 하면 된다. 하지만, 합성 단위가 단어 보다 작은 음절, 반음소 등이라면 연결 합성을 할 때 운율을 정교하게 예측하는 과정이 필요하다. 작은 합성 단위를 사용한 음성 합성에서 운율을 예측하기 위해서는 다양한 환경과 다양한 장르의 자연 발화를 가능한 한 많이 구축한 후, 훈련 과정을 통해 특정 문법 구조, 특정 환경, 특정 장르, 특정 화자에 따라 개별 분절음 등의 운율을 예측하는 모델을 구축한 후 새로운 문장이 입력되면 예측 모델을 활용해 입력된 문장에 맞는 최적의 운율을 예측해 주는 것이다. 운율 가운데 분절음의 길이를 예측하는 모델 중 하나를 예로 들면 아래와 같다(Klatt 1973; 1979; 1987).

$$D_j = K * (D_{inh} - D_{min}) + D_{min}$$

위의 모델은 대규모 음성 자료를 바탕으로 분절음의 길이를 예측하는 공식으로 D_j는 산출된 분절음의 길이, D_{inh}는 특정 분절음의 내재적 길이, D_{min}은 특정 분절음의 최소 길이(모음의 경우 내재적 길이의 약 45%), K는 환경에 따른 규칙 적용을 통해 결정된 변수이다. K는 다음의 환경을 기준으로 한 규칙을 통해 결정된 개별 값의 곱으로 표현된다.

규칙 1. 절 경계나 구두점 경계 앞의 묵음 구간
규칙 2. 어절 말 장음화
규칙 3. 어구 말 장음화
규칙 4. 단어 내 단음화

규칙 5. 다음절 단음화
규칙 6. 단어초가 아닌 자음의 단음화
규칙 7. 비강세 분절음의 단음화
규칙 8. 강조가 있는 모음의 장음화
규칙 9. 무성 자음앞 모음의 단음화
규칙 10. 자음연쇄의 단음화
규칙 11. 기식파열음으로 인한 강세 모음이나 공명음의 장음화

즉, 위의 공식과 규칙에 따라 훈련을 위해 구축되어 있는 음성 자료의 분절음 길이를 가장 잘 예측해 줄 수 있도록 개별 분절음의 D_{inh}와 D_{min}, K의 값을 찾고 그것을 통해 새로 입력되는 문장의 분절음 길이를 예측하는 모델이다. Klatt(1976)에서 제시된 모음 길이 예측 규칙을 예는 아래와 같다. 틀문장에서 발화된 단어의 모음 /æ/와 /ɪ/의 길이 예측 규칙이다.

D_{inh}는 유성파열음으로 끝나는 어말 단음절 단어에서 추출한다(예: “bag,” “big”)(Klatt 1976, 1217)

Phone	D_{inh}	D_{min} (D_{inh} × Ratio)	Ratio
/æ/	240	105	0.42
/ɪ/	160	65	0.42

규칙 1. 모음 뒤 파열음이 무성이면 모음의 길이를 45 msec 줄이시오.

$$D = D - 45$$

규칙 2. 모음이 어말이 아니면 길이를 35% 줄이고. 즉, 아래 공식에서 D_{min}은 내재적 모음 길이의 약 절반이기 때문에 K를 0.6으로 설정하시오.

$$D_j = K * (D_{inh} - D_{min}) + D_{min}$$

규칙 3. 비강세 모음의 K를 0.4로 설정해 길이를 줄이시오. 다음절 단어초의 비강세 음절에 오는 모음의 경우에는 K를 0.55로 설정하시오.

$$D_j = K * (D_{inh} - D_{min}) + D_{min}$$

규칙 4. 다음절 단어의 모든 음절을 15% 줄이시오. 즉 K값을 0.78로 설정하시오.

$$D_j = K * (D_{inh} - D_{min}) + D_{min}$$

오늘날에는 위의 공식을 좀 더 정교하게 만들어 다음과 같은 제곱합 모델을 통해 (Sums-of-Products model) 분절음의 길이를 예측하기도 한다(van Santen 1992).

$$DUR(d) = \sum_{i \in K} \prod_{j \in I_i} S_{i,j}(d_j)$$

위와 같은 제곱합 모델 이외에도 회귀분석을 활용한 CART(Classification and Regression Tree) 모델(Riley 1992)을 활용해 분절음의 길이뿐만 아니라 기본주파수를 예측하기도 한다.

이러한 규칙을 활용해 운율을 예측할 때는 언어학적으로 의미 있는 모든 규칙을 적용하는 것이 최선이겠지만, 기존의 훈련 자료를 가장 잘 설명하고 예측할 수 있는 최적의 규칙만을 활용하고 운율의 예측에 기여하지 못하는 규칙은 활용하지 않는다. 경우에 따라서는 운율의 예측에 기여하지 못하는 규칙을 모델에 활용하게 되면 음성 합성기의 성능을 오히려 떨어뜨리거나 합성의 속도를 느리게 하는 결과를 가져올 수도 있다.

다음으로 실제 기본 단위를 서로 연결 concatenation하는 과정이 필요한데 이 과정에서 기본 단위 간의 스펙트럼 신호 신호를 서로 부드럽게 연결해 주기 위해서 보간법 interpolation과 유연화 smoothing 기법이 필요하다. 합성 연결 부위에 있는 스펙트럼 신호간의 불일치가 크면 합성음이 울리거나 튀게 된다. 보간법과 유연화 기법으로는 PSOLA(Pitch Synchronous Overlap and Add; Carpentier & Moulines, 1990), FD-PSOLA(Frequency Domain PSOLA), TD-PSOLA(Time Domain PSOLA) 등이 많이 사용되고 있다. 이러한 기법은 기본 단위에 있는 원 음성의 음향적 특질을 크게 왜곡시키지 않으면서도 분절음이나 발화의 길이, 기본주파수, 강도 등을 유연하게 조절할 수 있다는 장점이 있다.
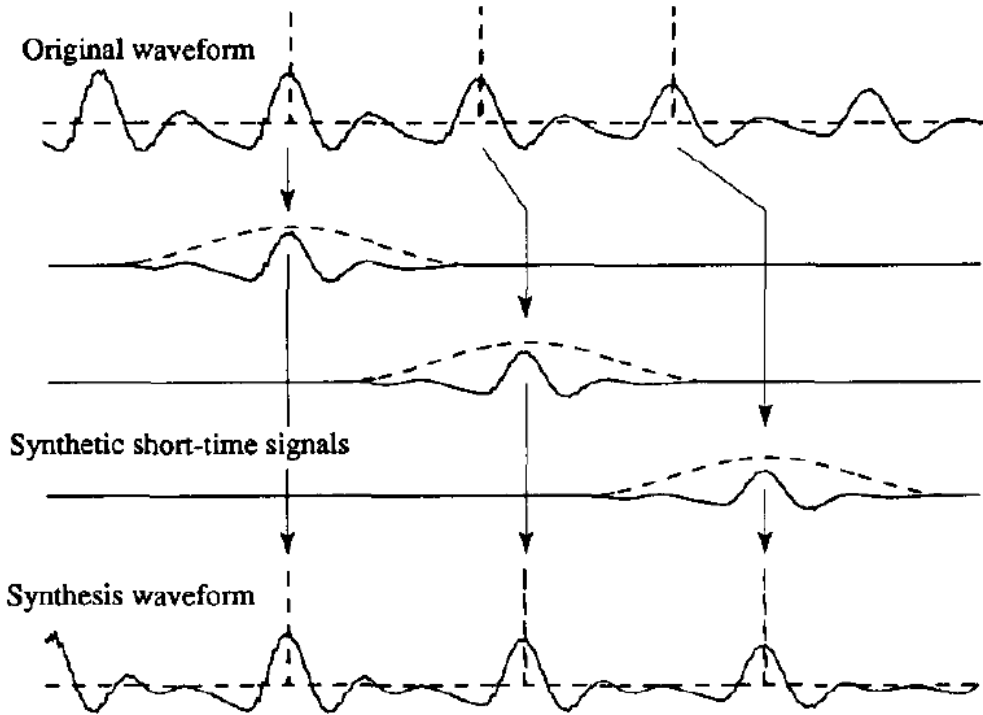
그림 8. TD-PSOLA 기법을 활용한 피치 조절(Moulines & Laroche 1995)

위 그림은 TD-PSOLA 기법을 사용해 발화의 피치를 조절하는 방식이다. 상단은 모음 /i/의 원래 원래 음파로 피치를 보여줄 수 있는 주기는 점선으로 표시되어 있다. 중간은 합성음에서 구현하고자 하는 인공적인 피치 주기로 피치를 원래의 음성보다 0.8배 낮게 설정하고자 한 것이다. 원래의 음성의 주기와 원하는 주기의 연결은 화살표로 표시되어 있다. 하단은 TD-PSOLA를 적용한 보간법을 통해 유연화가 이루어진 이후 피치가 변형된 합성음의 음파를 나타내고 있다.

V. 음성 합성의 실제: MBROLA 합성기를 이용한 미국 영어와 한국어 합성

최종 제품을 통한 음성 합성은 단순히 문자를 입력하고 산출된 합성음을 들어보는 과정이기 때문에 합성의 과정을 직접 체험해 볼 수 없다. 이 책에서는 본인이 직접 분절음을 선택하고, 길이 및 기본주파수 등의 운율을 조절해 볼 수 있는 MBROLA 음성합성기를 다루어 보기로 한다. MBROLA 음성합성기는 벨기에에 기반을 둔 Faculté Polytechnique de Mons의 TCTS Lab에서 주도한 MBROLA Project를 통해 개발된 것으로 반음소(diphone) 연결 합성을

이용하고 반음소간의 연결은 MBR-PSOLA(Multi-Band Re-Synthesis PSOLA) 알고리듬을 사용한다(Dutoit et al. 1996). 이 연결합성은 반음소를 예측된 운율에 따라 확장하거나 압축한 후, 이어진 반음소를 연결하고 최종적으로 억양을 부여하는 합성 방식이다. 반음소 간의 불일치는 스펙트럼 신호를 서로 부드럽게 연결해 주는 이 알고리듬 특유의 보간법 interpolation과 유연화 smoothing 기법을 통해 해소하고 있다. 현재 이 합성기법을 활용해 한국어를 포함해(Chung/Huckvale/Gim 1999) 음성합성이 가능한 30여개 언어의 반음소 음성 자료가 구축되어 있다.

이 합성기를 사용하기 위해서는 우선 홈페이지인 http://tcts.fpms.ac.be/synthesis/를 방문해 Download 페이지에서 MBROLA binary and voices 중에서 MBROLA binary는 자신의 컴퓨터 기종에 적합한 프로그램을 내려 받고, voices는 자신이 합성하기를 원하는 언어를 선택해 내려 받으면 된다. 본 책에서는 us1(미국영어)와 hn1(한국어)을 예로 들어 설명하기로 한다.

내려 받은 binary인 MBROLA Tools35를 설치한 후, us1과 hn1의 압축을 푼다. 설치된 프로그램의 제어판에서 us1과 hn1을 database로 선택할 수 있다. 두 언어 중 하나는 기본 database로 설정해 두어야 한다. 설치된 프로그램인 Mbroli를 열면 아래와 같은 창이 뜬다. 이 창에서는 반음소를 활용한 개별 소리와, 길이, 구간별 기본주파수 값을 입력할 수 있다.
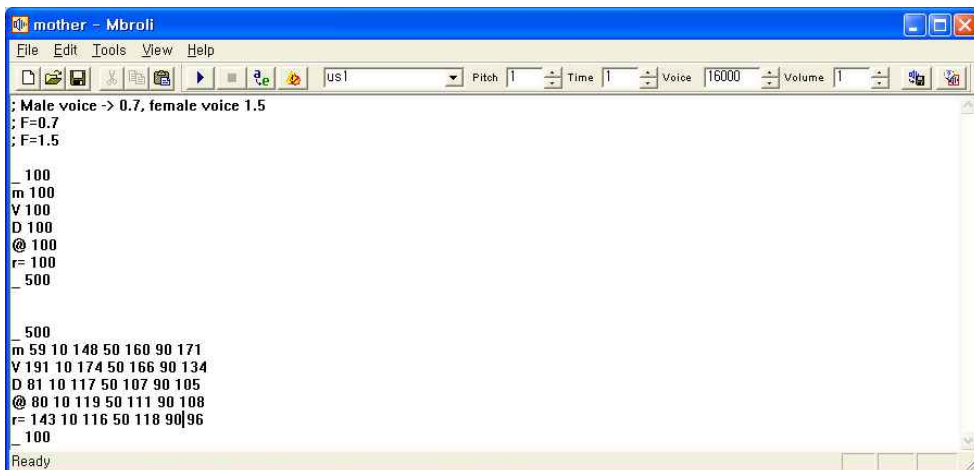
그림 9. MBROLA 합성기 미국영어 음성(us1)을 활용한 "mother"의 합성

위의 예에서 첫 번째 보기는 영어의 "mother"를 기본주파수 정보 없이 개별 음소의 길이를 100 밀리세컨드로 가정하고 주어진 반음소를 그대로 활용한 입력이고 아래는 실제로 발화된 발음의 길이와 구간별 기본주파수 값을 입력하여 합성해 본 것이다. 입력하는 발음기호는 특수 기호를 쓸 수 없기 때문에 ASCII 방식의 SAMPA(Speech Assessment Methods Phonetic

Alphabet; Wells, 1997) 전사 방식을 사용한다. SAMPA 전사기호는 언어별로 위 프로그램의 메뉴 가운데 Tools-Database Informations를 선택하면 확인할 수 있다. us1을 database로 한 입력 단어 “mother”의 경우 묵음 구간은 “_”, [m]은 “m”, [ʌ]는 “V”, [ð]는 “D”, schwa는 “@”, 음절말의 [ɹ]은 “r=”로 각각 전사할 수 있다. 입력 방식을 예로 들면 첫 분절음인 “m”의 경우 아래와 같다.

m 59 10 148 50 160 90 171

분절음 뒤의 첫 번째 값 “50”은 분절음의 길이를 밀리세컨드 단위로 입력한 것이고, “10 148”은 그 분절음의 길이가 10%가 진행된 구간의 기본주파수가 148 Hz라는 의미이고, 그 뒤의 숫자도 각각 50% 구간의 기본주파수가 160 Hz, 90% 구간의 기본주파수가 171 Hz라는 것을 나타낸다. 합성하는 방식은 일반인의 경우에는 본인의 소리를 직접 녹음하여 실제 분절음의 길이와 구간별 기본주파수를 측정한 후 위에서 제시된 음성 합성기에 입력하는 방법도 있겠지만, 음성 합성을 전문적으로 연구하는 사람들은 합성을 위한 문자열이 입력되면 입력된 문자열을 분석하는 과정을 거쳐 가장 적합한 반음소를 선택하고 위에서 제시된 운율 예측 모델링을 통해 분절음의 길이와 구간별 기본주파수 값을 자동으로 산출해 낸 후 합성에 활용하고 있다. 분절음을 바꾸거나 운율 정보를 변경하면 다른 합성음을 청취할 수 있다.

한국어 음성 합성을 원할 경우에는 본 연구자가 공동으로 개발한 database를 사용할 수 있다(Chung/Huckvale/Gim 1999). 제어판에서 기본 database를 hn1으로 선택하거나, 아래 그림과 같이 기본 database는 그대로 두고 필요할 때마다 프로그램의 첫 번째 선택 메뉴를 us1에서 hn1으로 수정해 주면 된다. 한국어도 전사 방식은 SAMPA 방식을 쓰고 있는데 전사 기호는 영어와 마찬가지로 database 선택을 완료한 후 메뉴에서 Tools-Database Informations를 선택하면 확인할 수 있다.
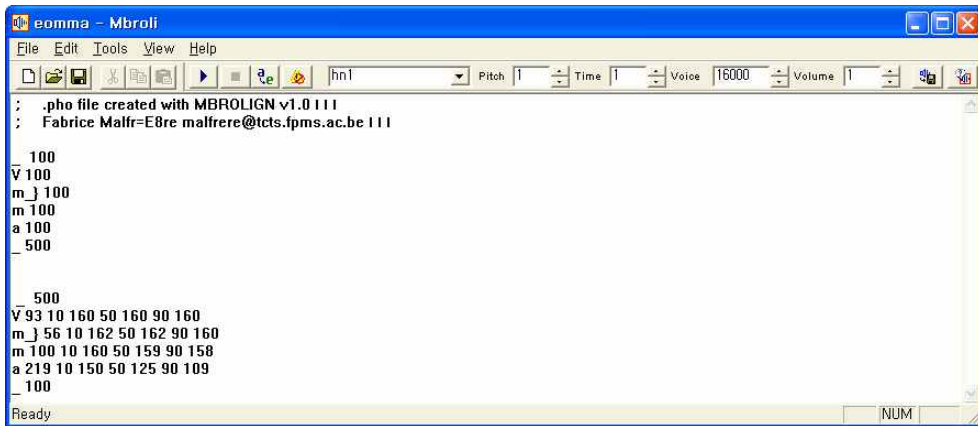
그림 10. MBROLA 합성기 한국어 음성(hn1)을 활용한 “엄마”의 합성

구체적인 합성 방식에 상관없이 단순히 문장을 입력한 후 합성음을 청취하고자 할 경우에는 http://www.talkforme.com/를 방문해 NextUp Talker를 내려 받아 실행해 볼 수 있지만 한 달간만 무료로 사용할 수 있다는 제한이 있다. 같은 곳을 방문하면 다양한 합성기를 통해 합성된 합성음을 들어볼 수도 있다. http://www.voiceware.co.kr/에서도 한국어를 비롯한 몇 개 언어의 합성음을 입력해 들어볼 수 있다.

VI. 음성 합성 기술의 교육적 적용

위에서 살펴본 음성 합성 기술은 실생활에서 활용이 확산되고 있는 추세이다. 가장 최근의 영문 OS가 설치되어 있다면 음성 인식과 합성 기술, 자연어 처리 기술, 인터넷 검색 기술이 접목되어 다음 그림과 같이 스마트폰 사용자와 스마트폰과의 양방향 대화도 가능하다. 양방향 대화의 원리는 일단 위에서 언급한 음성 인식 기술을 활용하여 스마트폰 사용자의 음성을 인식한 후 그것에 해당하는 명령을 수행하거나 사용자가 원하는 정보를 인터넷을 실시간으로 검색하여 화자에게 합성해 들려주는 것이다. 스마트폰은 사용자가 직접 손으로 입력해 검색하는 작업을 대신 수행하고 그것을 문자와 합성음으로 제공하는 것이다. 실시간으로 검색이 이루어지기 때문에 제공되는 정보는 역동성을 지니고 있다고 할 수 있다.

최근에는 자동 통역기가 개발되어 이 자동 통역기를 활용해 해외여행을 갔을 때 휴대가 간편한 스마트폰만 들고 가면 방문국의 언어에 대한 지식이 없더라도 스마트폰이 우리말을 인식한 후 그것을 해당 외국어로 변환하여 합성한 후 현지인에게 들려주고 현지인은 우리말을 할 필요 없이 현지어를 이야기 하면 현지어를 인식한 후 우리말로 변환하여 합성한 후 본인에게 들려주게 된다. 그 과정을 그림으로 도식화하면 아래 <그림 11>과 같다.
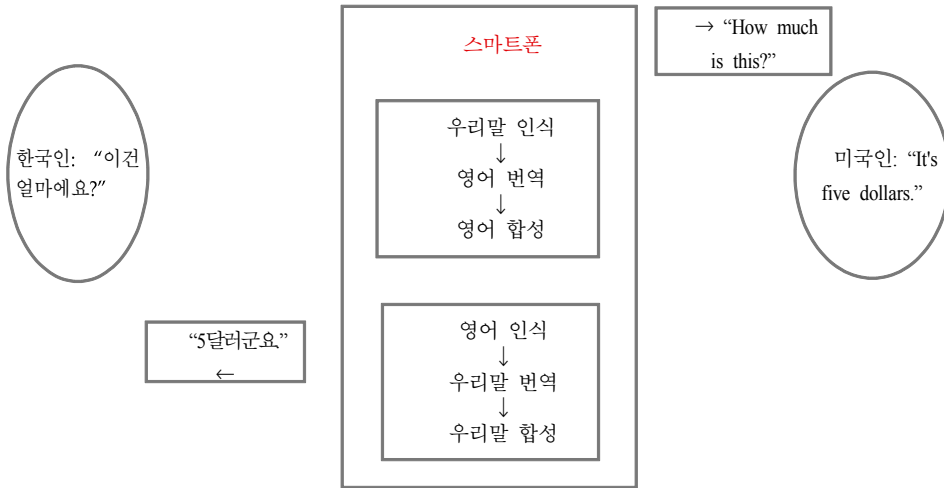
그림 11. 한영 자동 통역 시스템의 원리

과거에는 이러한 휴대용 기기를 활용하는 것이 불가능했기 때문에 음성 기술의 발달에도 불구하고 자동 통ㆍ번역기의 구현이 쉽지 않았지만 오늘날에는 이러한 기술을 구현할 수 기기의 소지가 일반화되면서 빠른 속도로 현실화되고 있다.

음성 기술은 영어의 문법 교육이나 발음 교육을 위해서도 활용될 수 있다. 가상의 대화 상대방과 의사소통을 하면서 자신의 표현을 수정해 볼 수도 있고, 자신의 발음을 녹음한 후 그 발음이 영어 원어민의 발음과 비교해 개별 분절음, 억양, 리듬이 얼마나 유사한지 무엇에 문제가 있는지 분석해 피드백을 받을 수도 있다.
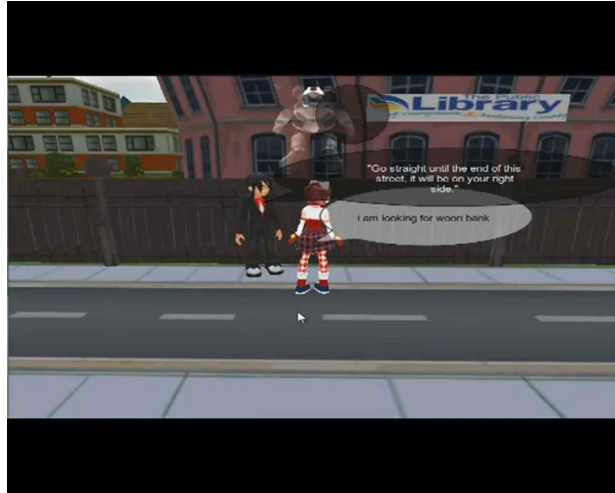
그림 12. POSTECH ISOFT의 문법 교정 프로그램
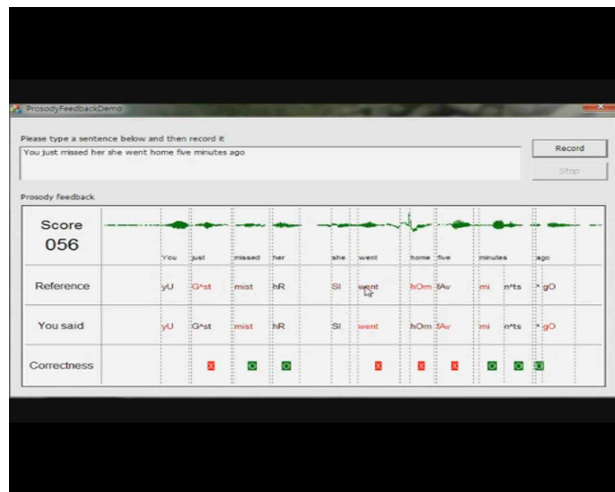
그림 13. POSTECH ISOFT의 리듬 교정 프로그램

VII. 음성 합성 기술의 미래

이 논문에서는 음성 합성의 원리와 연결합성의 응용 과정을 살펴보았다. 살펴본 것과 같이 음성학을 응용한 음성 합성 기술은 우리의 실생활에 다양하게 적용될 수 있고, 이를 위해 음성 기술은 눈부시게 발전해 왔다. 특히 영어 교육과 언어 교육을 위해서 원어민 없이도 책을 읽어 주는 등의 음성 합성 기술이 실제로 활용되고 있다. 하지만 아무리 자연스러운 합성음이

라고 하더라도 인간의 감정과 같은 것을 운율을 통해 완벽하게 구현하지 못하고 있기 때문에 담화적이거나 화용적인 연구가 더 진행될 필요가 있고, 더욱 다양한 실제 음성 자료를 구축해 운율을 예측할 수 있어야 할 것이다.

참고문헌

Allen, J./Hunnicutt, S./Klatt, D. H.(1987): From text to speech: The MITalk system. CUP: Cambridge, UK.

Chung, H./Huckvale, M./Gim, G.(1999): A new Korean speech synthesis system and temporal model. Proc. of 16th international conference on speech processing, 1, Seoul: The Acoustical Society of Korea & IEEE Korean Council, 203-208.

Coker, C. H.(1976): A model of articulatory dynamics and control. Proc. IEEE, 64, 452-459.

Cooper, F. S./Delattre, P. C./Liberman, A. M./Borst, J. M./Gerstman, L. J.(1952): Some experiments of the perception of synthetic speech sounds. J. Acoustic. Soc. Am., 24, 597-606.

Dudley, H./Riesz, R. R./Watkins, S. A.(1939): A synthetic speaker. J. Franklin Inst., 227, 739-764.

Dutoit, T./Pagel, V./Pierret, N./Bataille, F./van Der Vreken, O.(1996): The MBROLA project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes. Proc. ICSLP'96, Philadelphia, 3, 1393-1396.

Fant, G.(1953): Speech communication research. Ing. Vetenskaps Akad. Stockholm, Sweden, 24, 331-337.

Fant, G./Martonay, J.(1962): Speech synthesis, speech transmission laboratory, Royal institute of technology, Stockhom, Sweden QPSR, 2, 18-24.

Holmes, J. N.(1973): The influence of the glottal waveform on th naturalness of speech from a parallel formant synthesizer. IEEE Tran. Audio Electroacoust., AU-21, 298-305.

Kelly, J./Gerstman, L.(1961): An artificial talker driven from phonetic input. J. Acoust. Soc. Am. Suppl. 1, 33, S35.

Kelly, J./Gerstman, L.(1962): Digital computer synthesizes human speech. U.S. Patent 3, 158,685.

Klatt, D. H.(1973): Durational characteristics of prestressed word-initial consonant clusters in English. Research laboratory of electronics QPR, 108, MIT, Cambridge, MA, 253-260.

Klatt, D. H.(1976): Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. J. Acoust. Soc. Am., 59, 1208-1221.

Klatt, D. H.(1979): Synthesis by rule of segmental durations in English sentences. In: B. Lindblom & S. Öhman(Eds.), Frontiers of speech communication research(pp.

287-300):
Klatt, D. H.(1980): Software for a cascade/parallel formant synthesizer. J. Acoust. Soc. Am., 67, 971-995.
Klatt, D. H.(1987): Review of text-to-speech conversion for English. J. Acoustic. Soc. Am., 83(3), 737-793.
Mattingly, I. G.(1966): Synthesis by rule of prosodic features. Lang. Speech, 9, 1-13.
Moulines, E./Laroche, J.(1995): Non-parametric techniques for pitch-scale and time-scale modification of speech. Speech Communication, 16(2), 175-205.
Riley, M.(1992): Tree-based modelling of segmental durations. In Bailly, G./Benôit, C.(Eds.), Talking machines. theories, models and designs(265-273). Amsterdam, the Netherlands: North-Holland.
van Santen, J.(1992): Contextual effects on vowel duration. Speech Communication, 11(6), 513-546.
Wells, J. C.(1997): SAMPA computer readable phonetic alphabet. In D. Gibbon, R. Moore & R. Winski(Eds.), Handbook of standards and resources for spoken language systems(Part IV, Section B). Berlin and New York: Mouton de Gruyter.

Abstract

Cancatenative Synthesis in Text-to-Speech for Foreign Language Education

CHUNG, Hyunsong (Korea National Univ. of Educ.)

This paper introduces theoretical background of speech synthesis and the practical application of text-to-speech in concatenative synthesis. Concatenative synthesis has been being used to improve the naturalness of the synthesized speech. The process of the concatenative synthesis includes pre-processing, morphological tagging, syntactic parsing, prosodic parsing, letter-to-phoneme conversion, prosody modelling, and unit selection. The prosody of the synthesized units can be predicted by using rule-based prosody modelling, Sum-of-Products model, or Classification and Regression Tree (CART) model. The MBROLA application can be used to practice the synthesis by controlling the duration and pitch of individual sounds. The research on prosody prediction in different pragmatic contexts and in different emotional situations is required to use the technology for foreign language education.

Key words:
음성 합성, 연결 음성 합성, 문자-음성 변환, 발음 교육
Speech synthesis, Concatenative synthesis, Text-to-speech, Pronunciation teaching

논문투고일: 2015. 08. 06.
논문심사일: 2015. 08. 15.
게재확정일: 2015. 08. 25.